Published in: Proceedings of the 45th Annual Meeting of the SPIE,
San Diego, CA. July 30 -- August 4, 2000.

Random cascades on wavelet trees and their use in analyzing and modeling natural images

Martin J. Wainwright[a], Eero P. Simoncelli[b] and Alan S. Willsky[a]

[a]Laboratory for Information and Decision Systems; MIT, Cambridge, MA 02139
[b]Center for Neural Science & Courant Institute; New York University, New York, NY 10003

ABSTRACT

We develop a new class of non-Gaussian multiscale stochastic processes defined by random cascades on trees of wavelet or other multiresolution coefficients. These cascades reproduce a rich semi-parametric class of random variables known as Gaussian scale mixtures. We demonstrate that this model class can accurately capture the remarkably regular and non-Gaussian features of natural images in a parsimonious fashion, involving only a small set of parameters. In addition, this model structure leads to efficient algorithms for image processing. In particular, we develop a Newton-like algorithm for MAP estimation that exploits very fast algorithms for linear-Gaussian estimation on trees, and hence is efficient. On the basis of this MAP estimator, we develop and illustrate a denoising technique that is based on a global prior model, and preserves the structure of natural images (e.g., edges).

Keywords: wavelets, natural images, statistical models, denoising, random cascades

1. INTRODUCTION

Statistical image models underlie a variety of applications in image processing and low-level computer vision; accordingly, the past decade has witnessed an increasing amount of research devoted to developing stochastic models of images [e.g., 1–4]. Simultaneously, wavelet transforms and other multiresolution representations have profoundly influenced image processing and low-level computer vision [e.g. 5,6]. Moreover, multiscale theory has proven useful in modeling and synthesizing a variety of stochastic processes, including fractional Brownian motion [e.g., 7], as well as other Gaussian processes [e.g., 8,9]. The intersection of these three lines of research — statistical image models, multiscale representations, and multiscale modeling of stochastic processes — constitute the focus of this paper.

In this paper, we develop a mathematical framework for capturing the statistical properties of natural images, and show that it can be used as the consistent basis for a variety of image processing tasks. Our framework not only captures the key characteristics of natural images, but does so in a parsimonious manner that requires a small number of parameters. In particular, we define a new class of multiscale stochastic processes by "mixing" a white multiscale Gaussian process with a nonlinear function of a second Gaussian multiscale process (the premultiplier). These cascade models represent a significant variation on linear models defined on multiscale trees [e.g., 8], because the nonlinear mixing operation produces highly non-Gaussian statistics. Nonetheless, we are able to exploit the embedded linear-Gaussian structure to develop efficient and optimal algorithms for image processing. To be sure, a number of other researchers [e.g., 1–3,10] have studied and exploited the properties of natural images on which we focus here, and our approach has both some similarities and important differences with this earlier work. Later in the paper, we discuss these links both in image modeling (see Section 3.2), and in image denoising and coding (see Section 4.2). We refer the interested reader to [11] for a more complete description of our work.

1.1. The statistics of natural images

We begin by describing some important statistical properties of natural images. A first important characteristic is the fractal structure of natural images [e.g., 12,13]. Consistent with fractal behavior, a large body of empirical work has shown that the power spectrum of natural images obeys a $f^{-\gamma}$ law [e.g., 12,2]. This power spectrum is only a partial description, since natural images exhibit highly non-Gaussian behavior. Indeed, if natural images were Gaussian, then any linear operation (e.g., a wavelet transform) applied to the image ensemble should also yield Gaussian

MW supported by NSERC 1967 fellowship; AW and MW by AFOSR grant F49620-98-1-0349 and ONR grant N00014-91-J-1004; ES supported by NSF CAREER grant MIP-9796040.

statistics. However, when natural images are decomposed in a wavelet basis, the resulting marginal distributions tend to be highly non-Gaussian, with high kurtosis and extended heavy tails [see 12]. These properties are found for a wide range of filters and natural images. The heavy-tailed shape of these marginals have been modeled by a number of researchers [e.g., 6,3,4,14].

Another important feature of natural images is their approximate scale invariance or self-similarity. Intuitively, there should be no preferred scale in an ensemble of natural images, since (disregarding occlusion) the same scene is equally likely to be viewed from a range of distances. One manifestation of the scale invariance of natural images is their $f^{-\gamma}$ spectral characteristic. The marginal distributions of wavelet coefficients provide further support for approximate scale invariance. When they are renormalized by a factor depending geometrically on scale, the resulting histograms tend to coincide, as they should for a scale-invariant process [14].

Empirically, the coefficients of orthonormal wavelet decompositions of natural images tend to be roughly decorrelated [e.g., 3]. Some theoretical analysis supports this observation, in that the orthonormal wavelet transform provides a good approximation to the Karhunen-Loève basis of $f^{-\gamma}$ stochastic processes [15]. More recent work has shown that wavelet coefficients from natural images, despite being roughly uncorrelated, exhibit striking dependencies. In particular, they exhibit a striking *self-reinforcing* characteristic, in that if one wavelet coefficient is large in absolute value, then "nearby" coefficients (where nearness is measured in scale, position, or orientation) also are more likely to be large in absolute value. This self-reinforcing form of dependency is found for wavelet coefficients at nearby spatial positions, adjacent orientations and spatial scales, and over a wide range of natural images [16,17,3].

2. MATHEMATICAL PRELIMINARIES

2.1. Gaussian scale mixtures

In this section, we define and provide examples of a semi-parametric class of random variables known as Gaussian scale mixtures (GSMs). To begin, a GSM vector c is formed by taking the product of two independent random variables, namely a positive scalar random variable z known as the *multiplier* or *mixing variable*, and a Gaussian random vector u distributed as* $\mathcal{N}(0, \Lambda)$. With this notation, we have $c \stackrel{d}{=} \sqrt{z}u$, where $\stackrel{d}{=}$ denotes equality in distribution. The GSM variable c is specified by the choice of mixing variable. As a special case, the finite mixture of Gaussians corresponds to choosing the density of the mixing variable p_z to be a (discrete) probability mass function.

Conditions that characterize which random vectors can be represented as GSMs are given in [11,18]. The family of Gaussian scale mixtures includes several well-known families of random variables [see 19,4,11]. A classical example is the α-stable family which satisfies a generalized version of the central limit theorem [see 20]. The case $\alpha = 2$ corresponds to the familiar Gaussian, whereas variables with $0 < \alpha < 2$ have increasingly heavy tails as $\alpha \to 0^+$. A well-known example with heavy tails is the Cauchy distribution, which corresponds to $\alpha = 1$. The generalized Gaussian family (also known as the generalized Laplacian family) is described by a parameter $\alpha \in (0, 2]$. The choice $\alpha = 2$ again corresponds to a Gaussian, whereas $\alpha = 1$ is a symmetrized Laplacian. The generalized Gaussian family is often used to model the marginals of wavelet coefficients [e.g., 6,21,14,17], where the tail parameter when fit to empirical histograms is typically less than one. The symmetrized gamma family [19] is also important because it (like the α-stable) is infinitely divisible, a property emphasized in the context of natural images in [1].

In this paper, we will frequently exploit the fact that a large class of non-negative multipliers z can be generated by passing a Gaussian random variable x through a nonlinearity $h : \mathbb{R} \to \mathbb{R}^+$, thereby generating the multiplier in the form $z = h^2(x)$. We refer to the Gaussian quantity x as the *premultiplier* since it is the stochastic input to the nonlinearity h that generates the multiplier. Although it is often possible to determine explicitly the form of h corresponding to a given GSM, the precise form of multiplier may not be critical for the purpose of application. In this context, an advantage of the GSM framework is that it allows an arbitrary choice of the nonlinearity h, thereby permitting the use of GSMs which may confer a computational or analytical advantage. For application, we typically choose the nonlinearity from parameterized families of functions that generate random variables with ranges of behavior. For this paper, we focus on the family $\{(x^+)^\alpha \mid \alpha > 0\}$, which generates a class of variables with a range of tail behavior that is qualitatively similar to the symmetrized gamma and generalized Gaussian families.†

*The notation $x \sim \mathcal{N}(\mu, \Lambda)$ means that x is distributed as a Gaussian with mean μ and covariance Λ.

†Here the notation x^+ denotes the positive part of x, defined by $x^+ = x$ for $x \geq 0$ and 0 otherwise.

2.2. Multiscale stochastic processes

In this section, we introduce some of the basic concepts and results concerning linear multiscale models defined on trees. We limit our treatment to those aspects required for subsequent development; the reader is referred to other literature [e.g., 7–9] for further details of these models, and their application to a variety of 1-D and 2-D statistical inference problems.

The processes of interest to us are defined on a tree, as illustrated in Figure 1, where the nodes $s \in \mathcal{T}$ are organized into a series of scales, which we enumerate $m = 0, 1, \ldots, M$. At the coarsest scale $m = 0$ (the top of the tree) there is a single node $s = 0$, which we designate the *root node*. At the next finest scale $m = 1$ are q nodes, that correspond to the *children* of the root node. We specialize here to regular trees, so that each parent node has the same number of children (q). This procedure of moving from parent to child is then applied recursively, so that a node at scale $m < M$ gives birth to q children at the next scale $(m + 1)$. These children are indexed by $s\alpha_1, \ldots s\alpha_q$. Similarly, each node s at scale $m > 0$ has a unique parent $\bar{\gamma}s$ at scale $(m - 1)$.

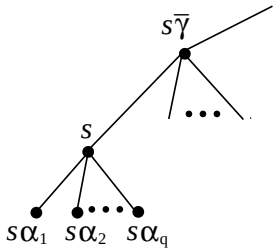


Figure 1: A segment of a q-adic tree, with the unique parent $s\bar{\gamma}$ and children $s\alpha_q, \ldots s\alpha_q$ corresponding to node s.

It should be noted that such trees arise naturally from multiresolution decompositions. For instance, a wavelet decomposition of a 1D signal generates a binary tree ($q = 2$), whereas decomposing an image will generate a quadtree ($q = 4$). To define a multiscale stochastic process, we assign to each node of the tree a random vector $x(s)$. The processes of interest to us are a particular class that are Markov with respect to the graph structure of the tree. A multiscale Markov tree process $x(s)$, $s \in \mathcal{T}$ has the property that for any two distinct nodes s, $t \in \mathcal{T}$, $x(s)$ and $x(t)$ are conditionally independent given $x(\tau)$ at any node τ on the unique path from s to t. Here we focus on Gaussian multiscale processes, specified by the distribution $x(0) \sim \mathcal{N}(0, P_x(0))$ at the root node, together with coarse-to-fine dynamics $x(s) = A(s)x(s\bar{\gamma}) + B(s)w(s)$ where the process noise is white[‡] on $\mathcal{T}$. Processes defined according to these dynamics in are called multiscale autoregressive (MAR) processes. It has been shown that the MAR framework can effectively model a wide range of Gaussian stochastic processes [7–9].

An additional benefit of the MAR framework is that it leads to extremely efficient algorithms for estimating the process $x(s)$ on the basis of noisy observations of the form $y(s) = C(s)x(s) + v(s)$ where $v(s)$ is a zero-mean white noise process with covariance $R(s)$. In particular, the optimal estimates of $x(s)$ at every node of the tree based on $\{y(s), s \in \mathcal{T}\}$ can be calculated very efficiently by a direct algorithm [8] with computational complexity of $\mathcal{O}(d^3 N)$ where d is the maximal dimension of $x(s)$ at any node, and N is the total number of nodes. This same algorithm also computes $P_e(s)$, the covariance of the error $[x(s) - \widehat{x}(s)]$ at each node $s \in \mathcal{T}$.

For notational reasons, it is useful to write down a vectorized form of the solution to the estimation problem. Let $\mathbf{x}$ be a vector formed by stacking the vectors $x(s)$ from each node $s \in \mathcal{T}$ in a fixed order, and define $\mathbf{y}$ analogously so that $\mathbf{y} = C\mathbf{x} + \mathbf{v}$ where C is a block diagonal matrix comprised of the $C(s)$ matrices, and $\mathbf{v} \sim \mathcal{N}(0, R)$ where R is the block diagonal matrix formed using the $R(s)$ matrices. The Bayes least-squares (BLS) and maximum a posteriori (MAP) estimates are identical in this case, and are given by

$$\widehat{\mathbf{x}} = P_{\mathbf{e}} C^T R^{-1} \mathbf{y} \qquad\qquad P_{\mathbf{e}} = \left[P_{\mathbf{x}}^{-1} + C^T R^{-1} C \right]^{-1} \tag{1}$$

where $P_{\mathbf{e}}$ is the covariance of the error $\mathbf{e} = \mathbf{x} - \widehat{\mathbf{x}}$. It is important to realize that for typical image processing problems (with several hundred thousand nodes), $\widehat{\mathbf{x}}$ and $P_{\mathbf{e}}$ are of extremely high dimension, and thus their computation as suggested by equation (1) is prohibitive. Instead, the fast tree algorithm solves the set of equations $P_{\mathbf{e}}^{-1} \widehat{\mathbf{x}} = C^T R^{-1} \mathbf{y}$ and simultaneously computes the diagonal blocks of $P_{\mathbf{e}}$, with the two pass procedure outlined previously.

[‡]Here we assume without loss of generality that means are zero, since it is straightforward to add in non-zero means.

3. RANDOM CASCADES ON WAVELET TREES

In this section, we introduce and develop a new type of multiscale stochastic process defined by random cascades on trees. As noted previously, naturally associated with a multiresolution decomposition like the wavelet transform is a tree of coefficients (a binary tree for 1D signals; a quadtree for images). Lying at each node is a random vector $c(s)$, which will be used to model a vector of d wavelet coefficients at the same scale and position, but different orientations. We model the wavelet vector $c(s)$ as a GSM of the form

$$c(s) \stackrel{d}{=} h(x(s)) \odot u(s) \tag{2}$$

where $x(s)$ and $u(s)$ are $d-$dimensional independent Gaussian random vectors. Here the nonlinearity h acts element-wise on the vector $x(s)$, and $\odot$ denotes element-wise multiplication of the two d-vectors. Here we assume that h has been appropriately normalized so that $\mathbb{E}[h^2(x_j(s))] = 1$ for $j = 1, \ldots d$ where $x_j(s)$ denotes the j^{th} element of the vector $x(s)$. Under this condition, $u(s)$ controls the variance of $c(s)$.

To specify a multiscale stochastic process, we need to define parent-to-child dynamics on the underlying state variables $x(s)$ and $u(s)$. We can express the covariance between $c(s)$ and its parent $c(s\bar{\gamma})$ as follows:

$$\text{cov}\left[c(s), c(s\bar{\gamma})\right] = \mathbb{E}\big[c(s)c^T(s\bar{\gamma})\big] = \mathbb{E}\Big\{ h(x(s))[h(x(s\bar{\gamma}))]^T \Big\} \odot \mathbb{E}\big[u(s)u^T(s\bar{\gamma})\big]$$

where we have used the independence of x and u. Since each element of $\mathbb{E}\{h(x(s))[h(x(s\bar{\gamma}))]^T\}$ is positive, this relation shows that the decorrelation of $c(s)$ and $c(s\bar{\gamma})$ is determined by the u process. Recall that for wavelet coefficients of natural images, the parent and child vectors are close to decorrelated. Therefore, to model wavelet coefficients of natural images, it is appropriate to choose $u(s)$ as a white noise process on the tree $\mathcal{T}$, uncorrelated from node to node. In contrast, the vector $x(s)$ must depend on its parent $x(s\bar{\gamma})$, in order to capture the strong property of local reinforcement in wavelet coefficients of natural images. Therefore, the GSM representation of equation (2) decomposes the wavelet vector $c(s)$ into two random components, one of which controls the correlation structure, while the other controls reinforcement among wavelet coefficients. We model the white noise process $u(s)$ as

$$u(s) = D(s)\zeta(s) \;, \qquad\qquad \zeta(s) \sim \mathcal{N}(0, I) \tag{3}$$

so that $D(s)$ controls any scale-to-scale variation (and hence the scaling law) for the process. To capture the dependency in the premultiplier process $x(s)$, we use a MAR model:

$$x(s) \quad = \quad Ax(s\bar{\gamma}) + Bw(s) \tag{4}$$

with $x(0) \sim \mathcal{N}(0, P_x(0))$ and $\zeta(s) \sim \mathcal{N}(0, I)$ at the root node. Although we specialize here to the stationary case of a MAR model (i.e., $A(s) \equiv A$ and $B(s) \equiv B$ for all nodes $s \in \mathcal{T}$), it is clear that GSM cascades with non-stationary MAR dynamics are also possible.

Equations (2), (3) and (4) together specify the random coefficients $c(s)$ of a multiresolution decomposition on a tree. Here each node s corresponds to a particular scale $m(s)$ and spatial location $p(s)$ in the image plane, and $c(s)$ is a random vector of wavelet coefficients for a set of different orientations at the same spatial location. These coefficients then define a random image via the inverse transform $\mathcal{I}(p_1, p_2) = \sum_{s\in\mathcal{T}} \sum_{i=1}^{d} c_i(s)\psi_{i;s}(p_1, p_2)$ where (p_1, p_2) is a point in the 2-D image plane, $c_i(s)$ is the i^{th} element of $c(s)$ (corresponding to the i^{th} orientation), and $\psi_{i;s}$ corresponds to the multiresolution basis element corresponding to orientation i, and centered at scale and position $(m(s), p(s))$.

3.1. Properties of GSM cascades

In other work [4,11], we have shown that GSM cascades capture the statistical behavior of wavelet coefficients from natural images. First of all, that the marginal densities of wavelet coefficients are well-fit by at least one GSM family — namely, the generalized Gaussian with tail exponent α used as a fitting parameter — is widely known [e.g., 6,21,14,17]. We have found that in addition, the symmetrized gamma family can also provide good fits to wavelet marginals [4]. Second, in contrast to local models, the global nature of our model (as defined by the tree structure) allows us to predict the joint distributions of any collection of coefficients. Empirically, it has been documented [22,4,14] that the joint distributions of wavelet coefficients exhibit a variety of shapes, ranging from

circular to concave star-shaped. We find that GSM cascades defined on trees account well for this range of shapes, as well as for the drop-off in dependence between pairs of coefficients as the spatial separation is increased [see 4]. Third of all, note that GSM tree processes are generated by a multiresolution transform discretized in scale, and therefore cannot be strictly self-similar. However, by choosing $x(s)$ to be stationary in scale and choosing $D(s) = 2^{-\gamma m(s)}$ in equation (3), we can ensure that they are dyadically self-similar. In particular, dyadic self-similarity of the random image $\mathcal{I}(p_1, p_2)$ means that $\mathcal{I}(p_1, p_2) \stackrel{d}{=} 2^{-k\gamma}\mathcal{I}\big(2^k(p_1, p_2)\big)$ for all integers k, where γ is a parameter. The parameter $\gamma > 0$ controls the drop-off in the power spectrum of the synthesized process [e.g., 7].

An attractive feature of the wavelet cascade models developed here is that they are specified by a rather small set of parameters: (a) the matrices $D(s)$ determine any scale-to-scale variation in the process, and hence the scaling law; (b) the choice of the nonlinearity h determines the form of the marginal distributions of wavelet coefficients, including tail behavior and kurtosis; and (c) the system matrices A and B determine the dependency of the underlying premultiplier process $x(s)$ from node to node. Here we investigate the effect of varying the nonlinearity h, as well as the system matrices. In particular, we simulate a one-dimensional cascade (i.e., the wavelet representation of a 1-D process) with the scaling $D(s) = 2^{-\gamma m(s)}$ with $\gamma = 1.5$; the nonlinearity $h(x) = (x^+)^\alpha$; and system matrices $A = \mu$ and $B = \sqrt{1-\mu^2}$ where the choices of the parameter α, and the scale-to-scale dependence μ were varied.

Figure 2 shows simulated random cascades for four combinations of the parameters (α, μ) using the 'Daub4' wavelet. The first three rows in each subfigure correspond to three scales of the wavelet pyramid, ranging from coarse to fine. The fourth row in each subfigure corresponds to the synthesized GSM process. First consider the effect of varying the parameter α. Note that the wavelet coefficients in cascades with $\alpha = 2$ (panels (c) and (d)) exhibit sparse behavior, in that a few outlying values tend to dominate. The wavelet coefficients of images also exhibit such sparsity, in that coefficients corresponding to edges and other discontinuities will tend to dominate. Of course, for both natural images and simulated cascades, this sparsity is a reflection of heavy tails in the marginal distributions. In contrast, wavelet coefficients in the cascades corresponding to $\alpha = 0.2$ (panels (a) and (b)) are distributed much more densely. In fact, histograms of these coefficients, as well as the behavior of the synthesized processes, are both quite close to Gaussian.

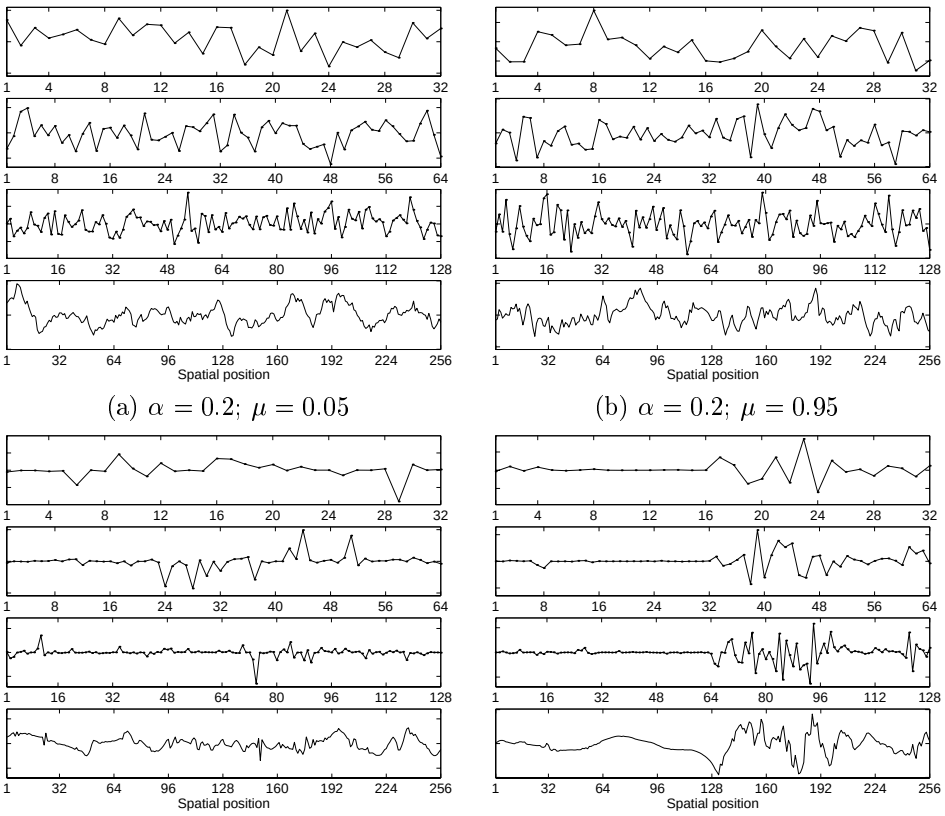
Varying the scale-to-scale dependence via the parameter μ also has a dramatic effect, particularly for the cascades with $\alpha = 2$. With $\mu = 0.05$ (panels (a) and (c)), coefficients from scale to scale are close to independent, so that high valued coefficients do not tend to form patterns through scale. In contrast, the high scale-to-scale dependence for the cascades with $\mu = 0.95$ generates trails of large magnitude coefficients through scale. One such trail is especially apparent in panel (d). These trails through the scale space of wavelet coefficients lead to a localized area of discontinuity and sharp variations in the synthesized process. In this respect, our GSM tree models constitute a precise analytical model for the cascade behavior exploited by successful image coders such as embedded zerotrees [e.g., 23].

3.2. Relation to previous work on image modeling

In this section, we discuss relations between GSM cascades on wavelet trees, and other approaches to image modeling. Simoncelli and colleagues [16,3,17] modeled the dependency between wavelet coefficients with a conditionally Gaussian model, where the variance of one wavelet coefficient depends on the absolute value or square of its neighbors. This type of model has proven useful in a variety of applications, including image coding, denoising, and texture synthesis. Our GSM cascades capture these same dependencies, but using an auxiliary multiplier variable that controls dependencies between coefficients. The multiplier variables are defined on a tree structure, thereby inducing a global probability distribution on the space of images, in contrast to the local model of Simoncelli et al.

Mumford and colleagues [e.g., 1,14] have defined and examined a number of properties of natural images, including (approximate) scale invariance, infinite divisibility of statistics, and highly kurtotic marginal distributions. As discussed previously, our GSM tree models satisfy an approximate form of scale invariance. Moreover, the marginal distributions of GSMs are highly kurtotic for many choices of multiplier variables, and particular choices ensure that the statistics will be infinitely-divisible (e.g., symmetrized gamma, α-stable.) As shown in previous work [4], our GSM tree models generate a range of behaviors in the joint contours of pairs of wavelet coefficients. Thus, our GSM cascades capture many of the properties emphasized by Mumford et al. in a parsimonious manner.

Our work is also related to the framework for non-Gaussian signal processing developed by Baraniuk and colleagues [24], and applied to image denoising in [10]. Their framework uses a hidden discrete-state process defined on a tree to capture dependencies between wavelet coefficients, which themselves are modeled as finite scale mixtures of



(a) $\alpha = 0.2$; $\mu = 0.05$ (b) $\alpha = 0.2$; $\mu = 0.95$

(c) $\alpha = 2$; $\mu = 0.05$ (d) $\alpha = 2$; $\mu = 0.95$

Figure 2. Simulated random cascades for various choices of the parameters. The first three rows of each panel correspond three scales of a wavelet transform, whereas the final row corresponds to the synthesized process. Heaviness of tails (and hence impulsiveness of the process) increases with the parameter α, whereas the parameter μ controls the scale-to-scale dependence.

Gaussians. It is important to note that for any finite mixture, the tails of wavelet marginal distributions will always drop off as $\exp(-c^2)$ like a Gaussian. In principle, by letting the number of multiplier states increase, one can push the Gaussian drop-off out further into the tails, thereby obtaining increasingly better models of coefficient marginal distributions. However, a very large number of states may be required to accurately model the tail behavior, as well as capture the high kurtosis around the origin. In terms of the parsimony of the model, there is a cost for increasing the number of states: namely, the number of parameters required to specify the model will increase as $\sim M^{2d}$, where M is the number of multiplier states, and d is the dimension of the multiplier vector at each node. In contrast, we have emphasized the use of infinite mixtures, which accurately capture both the highly non-Gaussian tail behavior and high kurtosis of wavelet marginal distributions with a small number of parameters.

4. ESTIMATION

We now turn to problems of estimation in GSM cascades on wavelet trees. Such problems involve using data or observations to make inferences about either the state (i.e., $x(s)$ and $u(s)$) of the GSM, or about unknown model parameters. In [11], we describe an algorithm for estimating the system matrices A and B of a GSM cascade. Here we limit ourselves to describing an algorithm for estimating the premultiplier $x(s)$, which is an important problem for a variety of applications in image processing (e.g., image coding and denoising). A significant benefit of the GSM framework is that conditioned on knowledge of the premultiplier, a GSM model reduces to a linear-Gaussian system, which can be analyzed by standard techniques.

4.1. Estimating the premultiplier

Here we address the problem of estimating the premultiplier $x(s)$ on the basis of noisy observations of the form

$$y(s) \quad = \quad h(x(s)) \odot u(s) + v(s) \tag{5}$$

where $v(s) \sim \mathcal{N}(0, R(s))$ is observation noise. An interesting feature of this problem is that unlike the case of linear observations in additive noise (see Section 2.2), the task of estimating $x(s)$ given noiseless observations (i.e., $R(s) \equiv 0$) is *not* trivial. Indeed, even in the absence of $v(s)$, the state $u(s)$ effectively acts as a multiplicative form of noise. With the noise $v(s)$ present, we have an estimation problem that is nonlinear, and includes both additive and multiplicative noise terms.

Given that we have a dynamical system defined on a tree, optimal estimation can, in principle, be performed by a two-pass algorithm, sweeping up and down the tree. For the linear-Gaussian case described in Section 2.2, computation of the optimal estimate (which is simultaneously the Bayes' least-squares (BLS) and maximum a posteriori (MAP) estimate) is particularly simple, involving the passing of conditional means and covariances only. In general, for nonlinear/non-Gaussian problems, however, not only are the BLS and MAP estimates different, but neither is easy to compute. However, the GSM models developed here have structure that can be exploited to produce an efficient and conceptually interesting algorithm for MAP estimation.

To set up the estimation problem, let $\mathbf{x}$ denote a vector formed by concatenating the state vectors $x(s)$ at each node, and define the vector $\mathbf{y}$ similarly. Recall that the computation of the MAP estimate involves the solution of an optimization problem $\widehat{\mathbf{x}}_{MAP} \triangleq \arg\min_{\mathbf{x}} \big[-\log p(\mathbf{x}|\mathbf{y}) \big]$. Herein we simply write $\widehat{\mathbf{x}}$ to mean this MAP estimate. At a global level, our algorithm is a Newton-type method applied to the objective function $f(\mathbf{x}) \triangleq -\log p(\mathbf{x}|\mathbf{y})$. That is, it entails generating a sequence $\{\mathbf{x}^n\}$ via the recursion

$$\mathbf{x}^{n+1} = \mathbf{x}^n + \alpha^n S^{-1}(\mathbf{x}^n) \nabla f(\mathbf{x}^n) \tag{6}$$

where the matrix $S(\mathbf{x}^n)$ is the Hessian of f, or some suitable approximation to it; and α^n is a stepsize parameter. This class of methods is attractive [see 25], because under suitable regularity conditions, not only is convergence to a local optimum guaranteed, but in addition the convergence rate is guaranteed to be quadratic. The disadvantage of such methods, in general, is that the computation of the descent direction $d^n \triangleq S^{-1}(\mathbf{x}^n) \nabla f(\mathbf{x}^n)$ may be extremely costly. This concern is especially valid in image processing applications, where the dimension of the matrix $S(\mathbf{x}^n)$ will be of the order 10^5 or higher.

One of the most important features of our model set-up is that the computation required for each step of equation (6) can indeed be performed efficiently. More precisely, the computation of the descent direction is equivalent to the solution of a *linear* MAR estimation problem, allowing the efficient algorithm of [8] described in Section 2.2 to be used for its computation. In order to demonstrate this equivalence, we begin by using Bayes' rule to express the objective function as $f(\mathbf{x}) \triangleq -\log p(\mathbf{x}|\mathbf{y}) = -\log p(\mathbf{y}|\mathbf{x}) - \log p(\mathbf{x}) + C$; where C is a constant that absorbs terms not depending on $\mathbf{x}$. The vector $\mathbf{x}$ is distributed as $\mathcal{N}(0, P_{\mathbf{x}})$, where the large covariance matrix $P_{\mathbf{x}}$ is defined by the system matrices A and B in equation (4). As a result, we can write the second term as $-\log p(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T P_{\mathbf{x}}^{-1} \mathbf{x} + C$ where C again absorbs terms not dependent on $\mathbf{x}$. Now observe that the data $y(s)$ at each node is conditionally independent of all other data given the state vector $\mathbf{x}$. As a result, the first term can be expressed as a sum of data terms over nodes so that we can write

$$f(\mathbf{x}) = -\sum_{s=1}^{N} \log p(y(s)|x(s)) + \frac{1}{2}\mathbf{x}^T P_{\mathbf{x}}^{-1} \mathbf{x} + C \tag{7}$$

From the representation of f given in equation (7), it can be seen that the Hessian of f has the form $\nabla^2 f(\mathbf{x}) = P_{\mathbf{x}}^{-1} + D(\mathbf{x})$ where $D(\mathbf{x})$ is a block diagonal matrix, with each block corresponding to a node s. With this form of the Hessian, the descent direction d^n is given by

$$d^n = \big[P_{\mathbf{x}}^{-1} + D(\mathbf{x}^n) \big]^{-1} \nabla f(\mathbf{x}^n) \tag{8}$$

We now compare the form of equation (8) to the form of a linear-Gaussian problem given in equation (1). It is clear that the two problems are equivalent with appropriate identification of data terms, observation matrix, and noise covariance. Further details of these identifications can be found in [11]. Note that the overall structure of this MAP

estimation algorithm is of a hybrid form. The Newton-like component involves an approximation of the objective function f that is performed *globally* on the entire graph at once. Local graphical structure is exploited within each iteration where the descent direction is computed by extremely efficient and direct algorithms for linear multiscale tree problems [8].

Another important characteristic of the GSM framework is that conditioning on the premultiplier $x(s)$ reduces the model to the linear-Gaussian case. If, indeed $x(s)$ were known exactly, we would have that $P_c(s) = H[x(s)]P_u(s)H[x(s)]$ where $P_u(s) = D(s)D^T(s)$ is the covariance of $u(s)$, and $H[x(s)] \triangleq \text{diag}\{h(x(s))\}$. This suggests a suboptimal estimate in which we replace $x(s)$ by $\widehat{x}(s)$ — namely:

$$\widehat{c}(s) = \widehat{P}_c(s)\big[\widehat{P}_c(s) + R(s)\big]^{-1} y(s) \tag{9}$$

where $\widehat{P}_c(s) = H[\widehat{x}(s)]P_u(s)H[\widehat{x}(s)]$. It is this form of wavelet estimator that we use in our application to image denoising in Section 5.

4.2. Relation to other denoising techniques

There are a number of interesting links between the GSM tree estimator developed here, and previous approaches to wavelet denoising. Here we briefly summarize these links; a more detailed exposition can be found in [11]. First of all, a large class of approaches to denoising are *pointwise*, so-called because they operate independently on each wavelet coefficient. The link to the GSM framework comes from the Bayesian perspective, in which many of these methods can be shown to be equivalent to MAP or Bayes least-square (BLS) estimation under a particular kind of GSM prior for the marginal distribution. For example, soft shrinkage [26], a widely studied form of pointwise estimate, is equivalent to a MAP estimate with a certain GSM prior — namely, a Laplacian or generalized Gaussian distribution with tail exponent $\alpha = 1$ [see 27,21]. It is shown in [17] that by varying the tail parameter α of a generalized Gaussian prior, it is possible to derive a full family of pointwise Bayes least-squares (BLS) estimators.

The GSM framework can also be related to the James-Stein estimator (JSE), a technique with an often controversial history [28]. The JSE applies to the problem of estimating the fixed mean c of a multivariate normal distribution from noisy observations $y = c + v$, where $v \sim \mathcal{N}(0, \sigma^2 I)$. The *empirical Bayesian* [see, e.g. 29] viewpoint links the JSE to our GSM framework. In the empirical Bayesian derivation of the JSE, the unknown mean c is decomposed into two parts as $c = \tau u$ where $u \sim \mathcal{N}(0, I)$, and τ is an unknown but fixed quantity. Interestingly, this corresponds to a particular type of Gaussian scale mixture. As with our GSM wavelet estimation scheme, the JSE proceeds by estimating τ, and then substituting this estimate into the usual linear-Gaussian formula. Further details of this link between the GSM framework and the JSE can be found in [11].

Although not always explicitly stated, many other approaches to image denoising and image coding rely on a GSM type decomposition. One approach is to model dependency between the variance of a subband coefficient and its neighbors directly, using a conditionally Gaussian model [16,3,17]. Other techniques involve modeling wavelet coefficients as a scale mixture of generalized Gaussians [e.g., 30–32], or scale mixture of Gaussians [e.g., 33,34]. Some models permit the variance parameter to assume only a discrete set of values [e.g., 32], whereas others allows a continuum of values. The latter models effectively correspond to infinite mixture models, similar to those emphasized in the current paper. A step common to all these techniques, whether for denoising or coding, is to estimate the multiplier or variance. Many approaches use an estimate motivated by maximum likelihood (ML), based on a local neighborhood of coefficients [31,32,17,34]. In such a ML framework, the variance parameter is viewed as an unknown but fixed quantity, without a prior distribution. These forms of estimator are thus very close to the James-Stein estimator discussed previously. Overall, the GSM tree framework presented in this paper represents an extension from ML to MAP estimation, and from local to global prior models. Our models allow an arbitrary choice of the prior on the multiplier, which is controlled by the nonlinearity h. The GSM tree algorithm computes the MAP estimate based on a global prior model on the full multiresolution representation. This global model, which incorporates the strong self-reinforcing properties among wavelet coefficients, is induced by the multiscale tree structure.

In the context of the underlying tree, our GSM cascade models are closely related to the non-Gaussian modeling framework of Baraniuk et al. [24,10]. In their models, a multiscale discrete-state multiplier process defined on a tree controls the dependency among wavelet coefficients, which are modeled as finite scale mixtures of Gaussians. For finite mixtures in which multiplier variable takes on discrete values, there exist direct recursive algorithms for computing the marginal distributions of the discrete multiplier states conditioned on the data. The BLS estimate of wavelet coefficients given noisy observations can be obtained by taking expectations over these marginal distributions [see

24]. However, the computational complexity of computing marginal distributions scales exponentially as $\sim M^d$, where M is the number of multiplier states and d is the dimension of the multiplier. In practice, therefore, both the number of states and dimension of the multiplier may be limited; for example, the denoising algorithm of [10] uses a low and high variance state ($M = 2$), and a scalar multiplier at each node ($d = 1$). A small number of multiplier states means that the models may not properly capture the non-Gaussian tail behavior and high kurtosis of wavelet marginals (see Section 3.2), whereas a low multiplier dimension will restrict the modeling of dependencies between orientations. In contrast, our GSM modeling framework emphasizes infinite scale mixtures of Gaussians. As we have illustrated, these infinite mixtures accurately capture the non-Gaussian tail behavior and high kurtosis of wavelet coefficients. Regardless of the particular GSM used, the complexity of our algorithm scales as $\sim d^3$, where d is the dimension of multiplier vector at each node.

5. ILLUSTRATIVE RESULTS

Here we illustrate the application of the GSM-tree framework to denoising natural images, using an overcomplete multiresolution decomposition described in [35] called the steerable pyramid. In all cases, we use a decomposition with four orientations, which corresponds to a state dimension of $d = 4$. Therefore, lying at each node of a quadtree are the two 4-vectors $x(s)$ and $u(s)$, which are used to model the 4-vector of wavelet coefficients $c(s)$. By the notation $c_j(s)$, we mean the coefficient at scale s and orientation j. We refer to a collection of all coefficients at the same scale and orientation (but different spatial positions) as a subband. Noisy observations of the wavelet coefficients are given by equation (5), where $R(s) = \sigma^2 I$.

Recall that the GSM-tree algorithm first computes the MAP estimate of the premultipliers $x(s)$, which it then uses to compute denoised wavelet coefficients via equation (9). We have experimented with different choices of the nonlinearity h, including the families $\{\exp(\alpha x \,|\, \alpha \geq 0\}$ and $\{(x^+)^\alpha | \alpha \geq 0\}$. As a Newton-like method, convergence of the algorithm tends to be rapid for sufficiently smooth (i.e., C^2) choices of this nonlinearity. Given the denoised multiresolution coefficients $c(s)$, the clean image is obtained by inverting the multiresolution decomposition.

We compare the denoising behavior of the GSM-tree algorithm to a number of other techniques. With the exception of one algorithm (MATLAB's adaptive filtering), all techniques are applied to the steerable pyramid decomposition, and involve an estimate of the subband variance. This estimate is given by $\sigma_c^2 = [\mathrm{var}(y(s)) - \sigma_n^2]^+$ where σ_n^2 is the variance of the noise in the subband (which can be computed directly from σ). All of the algorithms compared here are semi-blind, in that we assume that the noise variance σ^2 is known. The techniques to which we compare our algorithm here are:

(a) *Wiener subband technique:* for each subband, compute denoised coefficients as $\widehat{c}_j(s) = \sigma_c^2 \left[\sigma_c^2 + \sigma_n^2\right]^{-1} y_j(s)$ where σ_c^2 is the variance of the subband, and σ_n^2 is the noise variance in that subband.

(b) *Adaptive:* MATLAB's adaptive filtering routine called by *wiener.m*: it performs pixel-wise Wiener filtering with a variance computed from a local 5×5 neighborhood.

(c) *Soft thresholding:* For each subband, perform soft thresholding [26], where the threshold $t = \lambda \sigma_n^2/2$ is determined by the noise variance σ_n^2 and the scale parameter λ of a Laplacian distribution fit to the subband marginal.

We have applied these algorithms to a variety of natural images. In Figure 3, we depict representative results for the 256×256 "Einstein" image, using the nonlinearity $h(x) = (x^+)^5$. Shown in Table 1 are the SNR in decibels (dB) of the denoised images for all algorithms, based on original noisy images at four levels of SNR. For all levels of SNR, the GSM tree algorithm is superior to other techniques. Figure 3 depicts cropped denoised images for the "Einstein" image (a), on the basis of the noisy observations (SNR 4.80 dB) shown in (b). Panels (c), (d), (e) and (f) show the results of the Wiener subband denoising, MATLAB adaptive filtering, soft thresholding, and the GSM tree algorithm respectively.

Note that the the GSM estimator suppresses noise in regions where the multiplier $h(x(s))$ is of low amplitude, while simultaneously preserving peaks in high amplitude regions. In an *average sense*, it can be shown [11] to behave similarly to a form of shrinkage or soft thresholding [e.g., 26,17], in that it preferentially shrinks smaller observation values while modifying larger ones much less. Based on the discussion in Section 4.2, this is not surprising since many forms of thresholding, when interpreted in a Bayesian framework, correspond to a simple pointwise GSM model. Of

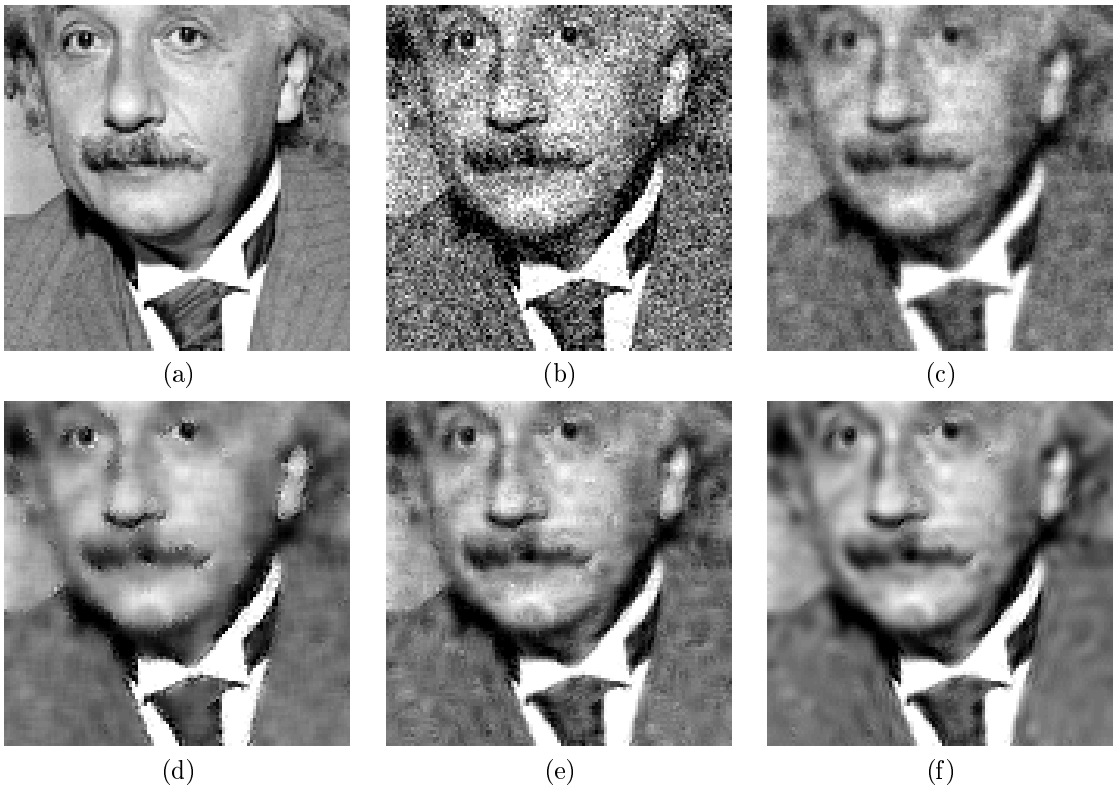


Figure 3. Cropped denoising results using a 4-orientation steerable pyramid. (a) Original image. (b) Noisy image (SNR 4.80 dB). (c) Wiener subband denoising. (d) MATLAB adaptive. (e) Soft thresholding. (f) GSM-tree algorithm.

course, it is important to emphasize that the GSM tree estimator is similar to thresholding only in this average sense. Thresholding is a deterministic operation applied pointwise to each coefficient, whereas our estimate of each coefficient is based all data, using a global prior model that incorporates the strong cascade dependencies among coefficients.

Noisy	Wiener subband	*wiener2.m*	Soft threshold	GSM Tree
1.59	9.28	10.19	10.11	10.54
4.80	10.61	11.86	11.47	12.31
9.02	12.58	13.37	13.24	14.68
13.06	14.96	14.23	15.41	16.83

Table 1. Denoising results (SNR in dB) for 256×256 Einstein image using a 4-orientation steerable pyramid. The original noisy SNR is given by $10 \log_{10}[\mathrm{var}(\mathcal{I})/\sigma^2]$, and the cleaned SNR is given by $10 \log_{10}[\mathrm{var}(\mathcal{I})/\mathrm{var}(\widehat{\mathcal{I}} - \mathcal{I})]$, where $\mathcal{I}$ and $\widehat{\mathcal{I}}$ denote the original and denoised images respectively.

6. CONCLUSION

In this paper, we have developed a semi-parametric class of non-Gaussian multiscale stochastic processes defined by random cascades on trees of multiresolution coefficients. This model class is rich enough to accurately capture the remarkably regular and non-Gaussian features of natural images. As we have pointed out, our methodology has strong intellectual ties to a variety of different image models and methods for image analysis, but our formalism differs in fundamental and, we believe, very important ways.

In particular, a first significant feature of our modeling framework is its parsimony: only a very small set of parameters are needed to specify a GSM wavelet cascade. This suggests that fitting such models from data is a far

better-posed problem than other approaches which require many more degrees of freedom to be specified. Secondly, the multiplicative structure of our models naturally and simply captures both the correlation structure of wavelet coefficients from natural images, as well as their dramatic non-Gaussian behavior. Our GSM framework makes explicit the structure exploited by previous approaches to image coding [e.g., 23,30,16]. Moreover, it allows pointwise estimators, such as shrinkage, to be extended to a statistically optimal joint estimator of wavelet coefficients based on a global prior model. In particular, the structure of GSM tree models leads to a method that uses fast linear algorithms as an engine for intermediate computations. The per iteration complexity of this algorithm is linear in the number of nodes, and cubic in the dimension of the wavelet vector at each node. Since convergence is typically rapid, the total complexity of the algorithm compares very favorably to other optimal estimation methods.

The work outlined in this paper represents a first step at developing a powerful statistical framework for modeling and analysis of natural images. While the characteristics outlined in the previous paragraphs suggest the promise of this framework, further work is required to realize this promise fully. First, previous empirical work [4] shows that a small set of multipliers is sufficient to describe a local neighborhood of wavelet coefficients. In contrast, models described in this paper use a number of multipliers equal to the number of wavelet coefficients. Estimating the order of the underlying multiplier process, though a challenging problem, is an important one in order to develop models of even more power. Second, in the current application to denoising, we have considered only fixed types of nonlinearity (e.g. $h(x) = (x^+)^\alpha$ for $\alpha > 0$). It is also possible to use a nonparametric form of this nonlinearity, which would allow the model to further adapt to the image under consideration, with no loss of efficiency. Finally, although tree models are very successful at capturing longer range dependencies, it is well-known that they may improperly model the dependency between nodes that correspond to nearby spatial positions in the original image but are widely separated in terms of tree distance. There are several ways to address the problem of these boundary artifacts. One approach is the so-called overlapping tree framework of [36], which retains the tree structure but uses nodes that overlap spatially. Another is to relax the requirement of a tree structure by introducing graphical connections between wavelet coefficients that are spatially close. Such graphical models with cycles raise other interesting algorithmic challenges in estimation, which we are currently addressing [37].

REFERENCES

1. D. Mumford and B. Gidas, "Stochastic models for generic images." Preprint available at http://www.dam.brown.edu/people/mumford.
2. D. L. Ruderman and W. Bialek, "Statistics of natural images: Scaling in the woods," *Phys. Rev. Letters* **73**(6), pp. 814–817, 1994.
3. E. P. Simoncelli, "Statistical models for images: Compression, restoration and synthesis," in *31st Asilomar Conf.*, pp. 673–678, IEEE Sig. Proc. Soc., November 1997.
4. M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," in *Neural Information Processing Systems 12*, vol. 12, pp. 855–861, December 1999. Paper available at http://ssg.mit.edu/group/mjwain/mjwain.shtml.
5. P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Comm.* **COM-31**, pp. 532–540, April 1983.
6. S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Pat. Anal. Mach. Intell* **11**, pp. 674–693, July 1989.
7. M. Daniel and A. Willsky, "The modeling and estimation of statistically self-similar processes in a multiresolution framework," *IEEE Transactions on Information Theory* **45**, pp. 955–970, April 1999.
8. K. Chou, A. Willsky, and R. Nikoukhah, "Multiscale systems, Kalman filters, and Riccati equations," *IEEE Trans. AC* **39**, pp. 479–492, March 1994.
9. M. Luettgen, W. Karl, A. Willsky, and R. Tenney, "Multiscale representations of Markov random fields," *IEEE Transactions on Signal Processing* **41**, pp. 3377–3396, December 1993.
10. J. Romberg, H. Choi, and R. Baraniuk, "Bayesian wavelet domain image modeling using hidden Markov trees," in *Proc. IEEE ICIP*, (Kobe, Japan), October 1999.
11. M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in modeling and analyzing natural images," *Applied Computational and Harmonic Analysis* , 2000. Special issue on wavelets; to appear.

12. D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am. A* **4**(12), pp. 2379–2394, 1987.
13. A. Turiel, G. Mato, N. Parga, and J. P. Nadal, "The self-similarity properties of natural images resemble those of turbulent flows," *Phys. Rev. Lett.* **80**, pp. 1098–1101, 1998.
14. J. Huang and D. Mumford, "Statistics of natural images and models," in *CVPR*, 1999.
15. A. H. Tewfik and M. Kim, "Correlation structure of the discrete wavelet coefficients of fractional Brownian motion," *IEEE Trans. Info. Theory* **38**, pp. 904–909, March 1992.
16. R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans Image Proc* **8**, pp. 1688–1701, December 1999.
17. E. P. Simoncelli, "Bayesian denoising of visual images in the wavelet domain," in *Bayesian Inference in Wavelet Based Models*, P. Müller and B. Vidakovic, eds., ch. 18, pp. 291–308, Springer-Verlag, New York, June 1999. Lecture Notes in Statistics, vol. 141.
18. D. Andrews and C. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society* **36**, pp. 99–102, 1974.
19. J. Rosinski, "On a class of infinitely divisible processes represented as mixtures of Gaussian processes," in *Stable processes and related topics*, S. Cambanis, G. Samorodnitsky, and M. Taqqu, eds., pp. 27–41, Birkhauser, Boston, 1991.
20. G. Samorodnitsky and M. Taqqu, *Stable non-Gaussian random processes: stochastic models with infinite variance*, Chapman and Hall, New York, 1994.
21. P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using a generalized Gaussian and complexity priors," *IEEE Trans. Info. Theory* **45**, pp. 909–919, April 1999.
22. C. Zetzsche, B. Wegmann, and E. Barth, "Nonlinear aspects of primary vision: Entropy reduction beyond decorrelation," in *Int'l Symp. Soc. for Info. Display*, vol. 24, pp. 933–936, 1993.
23. J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Transactions on Signal Processing* **41**, pp. 3445–3462, December 1993.
24. M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. on Signal Processing* **46**, pp. 886–902, April 1998.
25. D. Bertsekas, *Nonlinear programming*, Athena Scientific, Belmont, MA, 1995.
26. D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. on IT* **41**, pp. 613–627, 1995.
27. A. Chambolle, R. A. DeVore, N. Lee, and B. J. Lucier, "Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Proc.* **7**, pp. 319–335, March 1998.
28. W. James and C. Stein, "Estimation with quadratic loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 361–379, 1961.
29. B. Efron and C. N. Morris, "Limiting the risk of Bayes and empirical Bayes estimators – part II: The empirical Bayes case," *Journal of the American Statistical Association* **67**, pp. 130–139, 1972.
30. S. LoPresto, K. Ramchandran, and M. Orchard, "Image coding based on mixture modeling of wavelet coefficents and a fast estimation-quantization framework," in *Proc. Data Compression Conf.*, March 1997.
31. S. G. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," in *Proc. IEEE ICIP*, pp. 535–539, October 1998.
32. J. Liu and P. Moulin, "Image denoising based on scale-space mixture modeling of wavelet coefficients," in *Proc. IEEE ICIP*, vol. 1, pp. 386–390, October 1999.
33. M. Mihçak, I. Kozintsev, K. Ramchandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Sig. Proc. Let.* **6**, pp. 300–303, December 1999.
34. V. Strela, J. Portilla, and E. Simoncelli, "Image denoising using a local Gaussian scale mixture model in the wavelet domain," in *Proceedings of SPIE*, (San Diego, CA), July 2000.
35. E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Int'l Conf on Image Proc*, vol. III, pp. 444–447, IEEE Sig Proc Soc., (Washington, DC), October 1995.
36. W. Irving, P. Fieguth, and A. Willsky, "An overlapping tree approach to multiscale stochastic modeling and estimation," *IEEE Transactions on Image Processing* **6**, November 1997.
37. M. J. Wainwright, E. B. Sudderth, and A. S. Willsky, "Tree-based modeling and estimation of Gaussian processes on graphs with cycles." Submitted to NIPS, May 2000.